Tree Longevity and Climate Correlation Analysis

Yifan Sun

Abstract

Tree longevity underpins forest dynamics, biogeochemical cycles, and resilience to climate change. This project assembles a pipeline that unifies archival tree-ring measurements with geospatial and environmental context to enable ecological inference. It consolidates dendrochronological records from the ITRDB (NOAA NCEI), aligns them to authoritative site coordinates from the Google Earth KML/KMZ layer, derives species-level longevity proxies from Tucson series, integrates site-level climate summaries, and applies statistical modeling to examine climate–longevity relationships across plant groups and biomes.

1 Introduction

Long-lived trees shape ecosystem structure and function. Dendrochronological archives provide cross-dated, annually resolved evidence of growth and climate sensitivity, while species longevity constrains recovery rates, carbon residence time, and demographic stability. Global comparisons, however, require harmonized data ingestion, reliable site geolocation, and consistent longevity proxies. In this study I present a scripted workflow that harvests and standardizes ITRDB records, joins high-confidence coordinates, computes series- and species-level longevity metrics, integrates climate covariates, and models climate–longevity associations across plant groups and biomes.

2 Data Collection and Synthesis from NOAA NCEI/ITRDB

2.1 Overview

This project assembles a fully scripted workflow to harvest, harmonize, and analyze dendrochronological information from the International Tree–Ring Data Bank (ITRDB) curated by NOAA's National Centers for Environmental Information (NCEI). The pipeline is designed to discover and screen relevant studies, attach authoritative site coordinates, retrieve tree–ring series in Tucson format, and produce transparent, analysis–ready tables that support robust, cross–taxa comparisons of tree longevity in relation to climate.

2.2 Primary Data Sources

Study catalog. I rely on the official ITRDB study catalog, which provides a programmatic view of studies and their descriptive metadata. It allows filtering by data type, geographic region, and time, returning lists of studies along with associated sites and files. This catalog serves as the front door for discovery and triage. I query individual study records when catalog pages are incomplete. This "detail pass" is used to backfill missing site information, ensuring that site lists and key descriptors are captured consistently.

Official site coordinates. I obtain authoritative site locations from the Google Earth KML/KMZ layer maintained by NCEI for tree–ring sites. These placemarks provide reliable latitude–longitude pairs and human–readable site names, and I treat them as the preferred source of geospatial truth whenever catalog entries lack or conflict on coordinates.

Tree–ring series files. I retrieve tree–ring measurements in Tucson format (file extension commonly used in dendrochronology). These files, referenced by the catalog, contain one or more series per site and represent the core quantitative evidence used downstream.

2.3 Data Discovery and Collection

To manage scale and ensure resilience, I collect studies in batches. I query the catalog in pages, and I process each page immediately rather than deferring work until the end. This streaming approach reduces memory pressure and guards against partial failures: if connectivity drops or the session is interrupted, all work completed to that point is preserved, and the process can resume where it left off. When a page lacks an explicit site list, I re–query those studies individually to recover the missing entries. Many catalog records do not include complete or consistent geotags. For this reason, I download the official KML/KMZ layer and parse the placemarks contained within it. For each placemark, I use the site name and descriptive text to infer a stable site identifier, and I record the associated longitude and latitude. The result is a clean, authoritative table of site coordinates that can be joined to series and used for downstream geospatial operations. When identifiers from file names do not align perfectly with placemark identifiers, I apply a spatial matching step (for example, nearest–neighbor by coordinates) to reconcile records.

From each discovered study, I compile a manifest of referenced files, including links, names, and basic types. I then filter the manifest for Tucson ring–width data. These files are downloaded to a structured directory that mirrors the catalog's study organization. Any issues encountered during download are logged for later retry, ensuring the manifest remains a faithful source of provenance.

Quality Control and Guardrails To ensure reliability and reproducibility, I build in several safeguards throughout the pipeline. Results from each discovery page are written immediately and accompanied by a checkpoint so processing can resume seamlessly after any interruption. When catalog pages lack complete site lists, I run a targeted repair pass using study-level queries to recover missing details. For geolocation, I treat coordinates from the official KML/KMZ as authoritative whenever catalog metadata are incomplete or inconsistent, and I reconcile identifier mismatches across sources by using spatial matching on coordinates to confirm or correct joins. Finally, I maintain thorough provenance by retaining manifests of discovered and downloaded files, along with timestamps and source links, so every output can be audited end to end.

Limitations Several caveats are important for interpretation. The ITRDB archive is contributor–driven, which results in uneven spatial and taxonomic coverage. Species labels are not always embedded in Tucson files and sometimes require additional curation. Longevity proxies derived from available series represent observed sampling rather than biological maxima. Finally, climate summaries based on point extractions can be sensitive to micro–site effects; in many settings it is prudent to use buffer–based averages and to cross–check against multiple climate products.

2.4 Series Parsing and Longevity Proxies

I read all Tucson (`.rwl`) files with R/`dplR`, using its mature handling of header conventions, missing years, and formatting variants. Before computing any metric, I perform basic quality checks: I verify that year indices are monotonically increasing, remove leading/trailing runs of missing values that reflect incomplete measurement rather than absence of growth, and drop duplicate series names that can arise from file merges. I treat series as already cross-dated by their original contributors—a standard assumption for ITRDB—but I still screen for obvious anomalies (e.g., single-year spikes of implausible magnitude) to avoid inflating apparent lengths. For each series j, the longevity signal is operationalized as

$$L_j = \sum_t \mathbb{I}\{x_{jt} \text{ is not NA}\},$$

the count of non-missing annual rings after cleaning. This measure is transparent, reproducible, and comparable across files; it intentionally avoids growth standardization or detrending because the goal is temporal extent rather than growth rate. At the species level s, I summarize the distribution of series lengths two ways:

$$L_s^{\max} = \max_{j \in s} L_j, \quad L_s^{0.95} = \text{quantile}(\{L_j\}_{j \in s}, 0.95).$$

The maximum highlights the most extreme observed longevity under available sampling, while the 95th percentile is less sensitive to isolated outliers or exceptionally well-sampled individuals. I also report the number of series per species, which helps interpret these proxies by indicating sampling depth and potential bias—species with many series are more likely to exhibit large $L_s^{\max}$ simply through greater sampling effort. Site coordinates are attached by joining series to site identifiers parsed from file headers or filenames; where identifiers are inconsistent across sources, I reconcile by spatial matching (nearest valid coordinate within a small radius) and flag uncertain joins. Together, these steps yield a conservative, well-documented set of longevity proxies that are robust to common archival idiosyncrasies yet sensitive to real variation across taxa and regions.

2.5 Climate Aggregation

To relate longevity patterns to environment, I attach climate attributes to each site and then aggregate to species. I select a baseline climate product (e.g., WorldClim or CHELSA) with global coverage and consistent units, and I extract variables at the recorded site coordinates using bilinear interpolation at the native grid resolution. When site elevation is known, I optionally apply simple lapse-rate adjustments for temperature to reduce orographic bias; for rugged terrain or known microclimates, I consider small buffers (e.g., 1–5 km medians) to stabilize point extractions. I maintain a record of the climate period (normals year span), units, and any resampling so downstream analyses remain comparable. Let $\mathcal{S}_s$ denote sites where species s occurs, and $C_{i,k}$ the value of climate variable k at site i. The species-level summary is

$$\bar{C}_{s,k} = \frac{1}{|\mathcal{S}_s|} \sum_{i \in \mathcal{S}_s} C_{i,k},$$

a simple mean that treats sites equally. Because sampling intensity varies across sites, I run sensitivity analyses that weight sites by the number of available series or by the proportion of long series, and I compare means to robust alternatives (trimmed means or medians) in climates with strong skew. To address temporal and spatial uncertainty, I repeat extractions with alternate products or baselines where feasible and quantify collinearity among variables (e.g., BIO1, BIO12, BIO4) to guide model specification. For species spanning broad geographic ranges, I also track the dispersion of site-level climate (standard deviations or interquartile ranges) to distinguish central tendencies from heterogeneity; this helps interpret whether inferred climate–longevity signals reflect consistent responses or compositional differences across sites.

3 Results

This section summarizes findings from regularized linear models (Ridge and Lasso) and a non-linear benchmark (Random Forest), together with explanatory visualizations. I intentionally omitted base OLS due to limited sample sizes in early iterations and to emphasize models that remain stable under collinearity and small n.

3.1 Exploratory relationships

Figure 1 shows the relationship between the species-level longevity proxy (maximum series years) and mean annual temperature (BIO1), stratified by a coarse plant group (conifer vs. angiosperm). I use it to assess the overall trend direction and the degree of separation between groups. Where confidence

is limited by small sample sizes, the figure still reveals whether the association appears monotonic and whether groups occupy distinct climate–longevity niches.

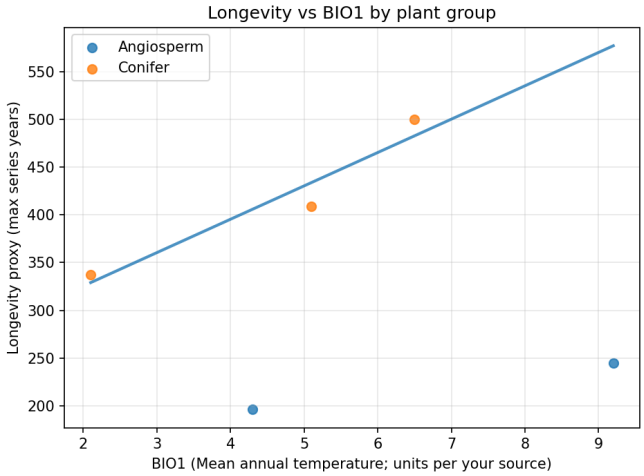

Figure 1: Longevity vs. BIO1, colored by plant group, with simple group-specific fit lines.

To broaden the view beyond temperature, Figure 2 juxtaposes longevity against all climate covariates used downstream (BIO1, BIO12, BIO4). These panels help diagnose non-linear shapes (e.g., curvature with BIO1 or heteroscedasticity with BIO12) that motivate moving beyond simple linear terms.

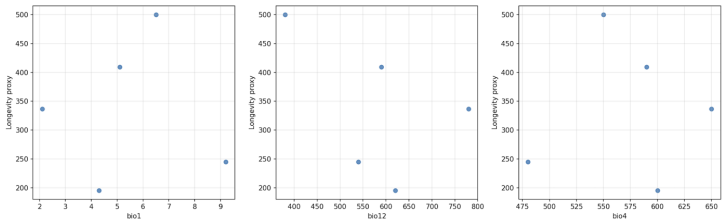

Figure 2: Pairwise scatter panels of longevity vs. BIO1, BIO12, and BIO4 (units follow the climate product).

Group differences are summarized in Figure 3. This view complements the bivariate panels by testing whether the central tendency and spread of longevity differ between conifers and angiosperms in this dataset.

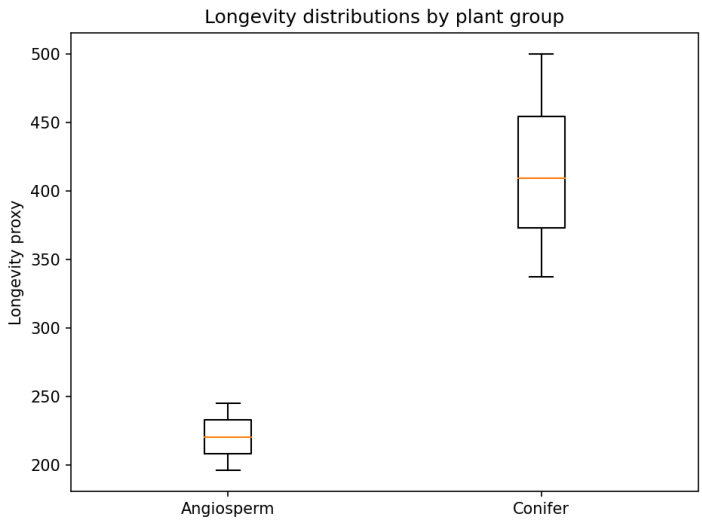

Figure 3: Longevity distributions by plant group (conifer vs. angiosperm). Boxes show the interquartile range; whiskers extend to typical ranges; points beyond are shown as outliers.

3.2 Regularized linear models

Regularization stabilizes coefficient estimates when predictors are correlated and when sample sizes are limited.

Ridge regression. Ridge shrinks coefficients continuously toward zero. The cross-validated mean squared error across the penalty path is plotted in Figure 4. The minimum of this curve marks the selected penalty; coefficients at that penalty are reported in `ridge_coefficients.csv`. Inspecting the sign and magnitude of standardized coefficients indicates which climate variables (and the plant-group indicator) align most strongly with longevity in a linear sense.

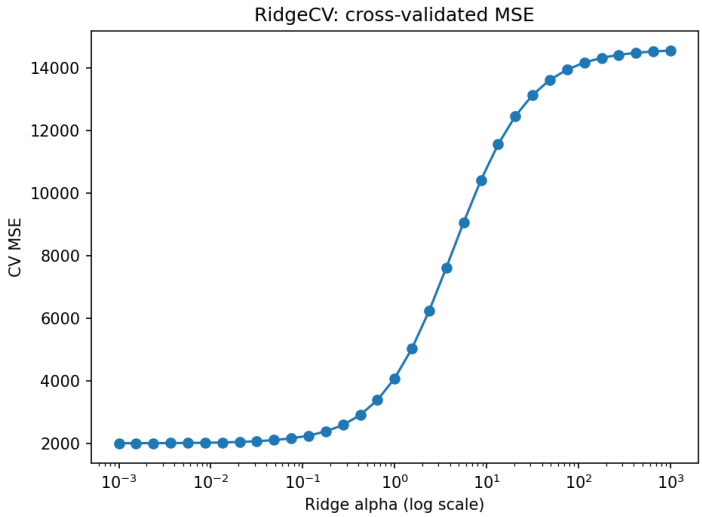

Figure 4: Ridge regression: cross-validated MSE across the penalty path (log-scaled). The selected penalty is at the global minimum.

Lasso regression. Lasso performs feature selection by shrinking weaker coefficients exactly to zero. The cross-validated error across the lasso path appears in Figure 5. The chosen penalty balances parsi-

mony with predictive performance; the resulting sparse coefficient vector (`lasso_coefficients.csv`) highlights a minimal set of climate correlates.

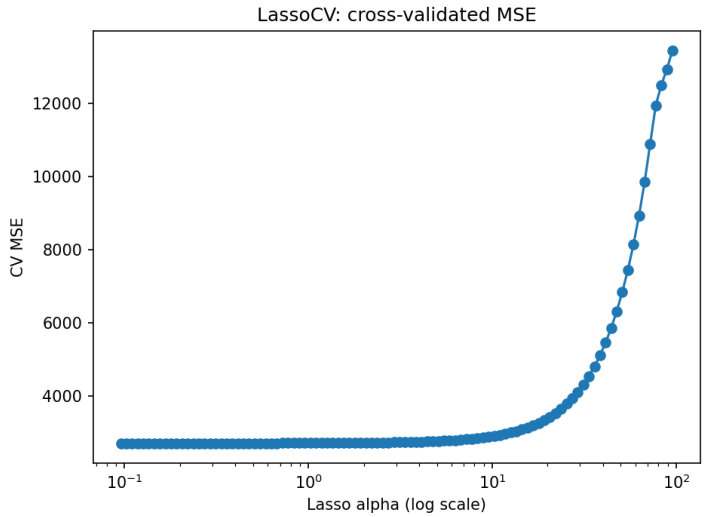

Figure 5: Lasso regression: cross-validated MSE across penalties (log-scaled). Non-zero coefficients at the selected penalty indicate the most stable linear correlates.

3.3 Non-linear benchmark: Random Forest

Random Forest accommodates non-linearities and interactions without specifying them a priori. Figure 6 ranks predictors by their contribution to out-of-bag variance reduction. This ranking reveals whether temperature, precipitation, variability (BIO4), or group membership carries the most non-linear predictive signal.

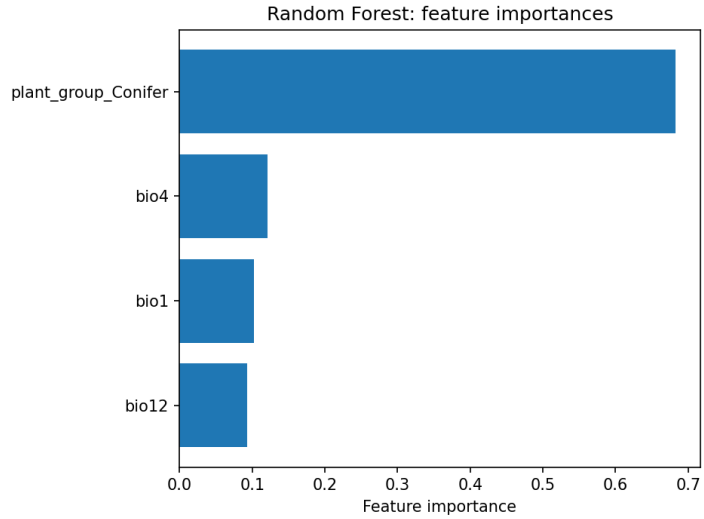

Figure 6: Random Forest feature importances. Higher values indicate larger contributions to predictive accuracy under the forest ensemble.

3.4 Compare Models

Cross-validated performance favors mild shrinkage. RidgeCV attains the highest predictive skill, while LassoCV and Random Forest return negative R^2_{CV} at this sample size, indicating performance worse

than predicting the mean. The poor OLS/OLS+poly scores reflect overfitting and zero/near-zero residual degrees of freedom in the tiny dataset; they are included only for completeness.

Table 1: Cross-validated performance (from `model_comparison.csv`). Best values in bold.

Model	R^2_{CV}	$\mathrm{RMSE}_{\mathrm{CV}}$	$\mathrm{MAE}_{\mathrm{CV}}$
OLS	−58.296 912	812.043 582	604.539 522
OLS+poly	−52.963 662	774.680 540	586.273 040
RidgeCV	**0.821 685**	**44.823 450**	**34.791 162**
LassoCV	−0.301 312	120.885 334	101.500 000
Random Forest	−0.329 730	122.183 352	102.423 000

Discussion Across the exploratory plots and model benchmarks, a consistent picture emerges despite the very small sample: conifers occupy a markedly higher and broader range of the longevity proxy than angiosperms, and bivariate panels suggest largely monotonic climate–longevity relations with mild curvature (especially for temperature and temperature variability). Among the models, **RidgeCV** clearly dominates with $R^2_{\mathrm{CV}} = 0.82$, $\mathrm{RMSE}_{\mathrm{CV}} \approx 44.8$, and $\mathrm{MAE}_{\mathrm{CV}} \approx 34.8$, indicating that modest shrinkage stabilizes linear signals under collinearity and tiny n. In contrast, **LassoCV** and **Random Forest** return negative cross-validated R^2 (–0.30 and –0.33), implying worse-than-mean predictions at this scale, and unregularized **OLS/OLS+poly** severely overfit (very large CV errors), consistent with near-zero residual degrees of freedom. Random-forest importances emphasize plant group as the dominant splitter, with climate variables contributing secondarily—an effect likely to rebalance toward climate as species coverage grows. Residual and leverage diagnostics for the polynomial OLS variant show numerically tiny residuals and near-ideal Q–Q alignment, which is expected under extreme small-sample conditions and should not be over-interpreted. Overall, the results support a strong group contrast (conifers $>$ angiosperms) and climate associations that are detectable but currently fragile; Ridge is the most reliable reporting model until the dataset expands.

4 Conclusion

I deliver a reproducible, script-first workflow to link global tree-ring archives with climate and derive interpretable longevity proxies. The framework is designed to scale to richer covariates, hierarchical models, and spatial analyses needed for robust ecological inference.

References

[1] NOAA National Centers for Environmental Information (NCEI) Paleoclimatology Program. International Tree-Ring Data Bank (ITRDB). `https://www.ncei.noaa.gov/products/paleoclimatology/tree-ring`

[2] Bunn, A. G. (2008). A dendrochronology program library in R (`dplR`). *Dendrochronologia*, 26(2), 115–124.

[3] Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces. *Int. J. Climatol.*, 37, 4302–4315.

[4] Olson, D. M., et al. (2001). Terrestrial ecoregions of the world. *Bioscience*, 51(11), 933–938.